

Optimal Shape-Gain Quantization for Multiuser MIMO Systems with Linear Precoding

Muhammad N. Islam, *Student Member, IEEE*, Raviraj Adve, *Senior Member, IEEE*, Behrouz Khoshnevis, *Member, IEEE*

Abstract

This paper studies the optimal bit allocation for shape-gain vector quantization of wireless channels in multiuser (MU) multiple-input multiple-output (MIMO) downlink systems based on linear precoding. Our design minimizes the mean squared-error between the original and quantized channels through optimal bit allocation across shape (direction) and gain (magnitude) for a fixed feedback overhead per user. This is shown to significantly reduce the quantization error, which in turn, decreases the MU interference. This paper makes three main contributions: first, we focus on channel gain quantization and derive the quantization distortion, based on a Euclidean distance measure, corresponding to singular values of a MIMO channel. Second, we show that the Euclidean distance-based distortion of a unit norm complex channel, due to shape quantization, is proportional to $2^{-\frac{2B_s}{2M-1}}$, where, B_s is the number of shape quantization bits and M is the number of transmit antennas. Finally, we show that for channels in complex space and allowing for a large feedback overhead, the number of direction quantization bits should be approximately $(2M - 1)$ times the number of channel magnitude quantization bits.

Index Terms

MIMO Broadcast Channels, Limited Feedback of CSI, Optimal Bit Allocation, Sum Mean Squared Error, Shape-Gain Product Quantization.

I. INTRODUCTION

The availability of channel state information (CSI) at the transmitter significantly improves the performance of multiuser (MU) multiple-input multiple-output (MIMO) systems [1]–[3].

Muhammad N. Islam is with WINLAB, Rutgers - State University of New Jersey, New Jersey, USA. email: mnislam@winlab.rutgers.edu. Raviraj Adve and Behrouz Khoshnevis are with Electrical & Computer Engineering, University of Toronto, Toronto, Canada. email: rsadve@comm.utoronto.ca, bkhoshnevis@comm.utoronto.ca

Specifically, CSI is essential for effective communications in the MU downlink. In frequency division duplex systems, in order to provide the base station (BS) with CSI, the receivers need to quantize the CSI and feed the quantized information back to the BS. Clearly this feedback is an overhead to the system and, therefore, must be limited to an acceptable level.

This paper focuses on limited-feedback MU MIMO systems, where a single BS communicates with multiple receivers and each user can potentially receive multiple data streams. Specifically, we restrict our analysis to systems using linear precoding [4]. Our main goal is to present an efficient quantization scheme for these systems. For this purpose, we use sum mean squared error across all data streams as the design objective and focus on quantization issues by assuming perfect channel estimation at the receiver (user) side and a noiseless, delay-free, feedback link to the BS.

This work is mainly motivated by the fact that the performance of limited-feedback MU-MIMO systems is very sensitive to the quality of the CSI available at the BS. Without accurate CSI, the quantization error common performance measures saturate in the high signal-to-noise ratio (SNR) regime because the BS cannot completely pre-cancel the multi-user interference [3], [5], [6]. It is therefore essential to design a limited feedback system such that the CSI quantization error is minimized. The optimal design of channel quantization in MU-MIMO systems is, therefore, the core objective of this paper.

Most of the works in limited feedback literature focus on either channel magnitude quantization [7]–[9] or direction quantization [1], [3], [10], [11] but not both. However, in MIMO systems, optimizing the precoder at the transmitter depends on both the channel magnitude (also known as the channel gain) and the phase of individual channel entries (the channel direction or shape). The authors of [12] specifically showed that one needs both channel gain and quantized direction information to achieve multi-user diversity gain. However, the gain information was assumed perfect in [12]. In general, joint vector quantization (VQ) of channel magnitude and direction is a very complex task [13]. To reduce the design complexity, the authors of [14] investigate independent quantization of the channel gain and shape and develop optimal bit allocation across gain and shape of real channel vectors using spherical codes. The authors of [15], [16] also use such a product codebook and solve for the optimal bit allocation to minimize the average transmit power with quality-of-service constraints. Such a structure has several practical advantages and provides an analytically tractable framework to optimize the limited feedback [14], [15], [17].

We adopt a similar approach where channel gain and shape are independently quantized.

The optimal quantizer depends on the transmission scheme and performance measure used. Due to its simplicity and efficiency, we adopt an eigen-based combining (EBC) approach to precode the data [1], [2]. We study quantization of CSI to minimize the sum mean squared error (SMSE) over all data streams received, a popular measure in the MU MIMO downlink [4], [18]. As shown in our earlier work [19], there is a one-to-one relationship between the SMSE objective and the variance of the quantization error. The current work, therefore, focuses on optimizing the bit allocation across shape and gain given a budget for feedback overhead per user. Once this bit allocation is optimized, one can use our earlier work in [19] for designing the limited-feedback system. To the best of our knowledge, the problem of optimal bit allocation in shape-gain vector quantization to minimize the SMSE of a multiuser MIMO system has not been investigated before.

This paper makes two key contributions:

- 1) We show that the quantization distortion of a uniformly distributed unit-norm vector in \mathbb{C}^M is upper bounded by: $K_s \times 2^{-\frac{2B_s}{2M-1}}$, where M is the total number of transmit antennas, B_s is the number of shape quantization bits and K_s is a constant that does not depend on B_s .
- 2) We also show that, for channels in complex space, the optimal number of channel direction quantization bits should be approximately $(2M - 1)$ times the optimal number of channel magnitude quantization bits.

Numerical simulations suggest that the proposed bit allocation laws provide a substantial improvement over full shape quantization or full gain quantization in the SMSE and bit error rate (BER) performance of a multiuser MIMO system.

Notation: Lower case letters denote scalar values while lower case bold face letters represent column vectors. Upper case boldface letters denote matrices. The superscripts $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose operators respectively. $\text{tr}[\cdot]$ denotes the trace operator. \mathbf{I} is reserved for the identity matrix whereas $\mathbf{1}$ represents the column vector with all one entries. $\text{diag}(a_1, \dots, a_n)$ denotes the diagonal matrix with diagonal elements a_1, \dots, a_n ; whereas $\text{diag}(\mathbf{A}_1, \dots, \mathbf{A}_n)$ represents the block diagonal matrix with the matrices $\mathbf{A}_1, \dots, \mathbf{A}_n$ on its main diagonal. $\|\cdot\|_1$ denotes the L_1 norm of the vector. $E(\cdot)$ and $S(\cdot)$ denote statistical expectation and surface area respectively.

The remainder of the paper is organized as follows. Section II outlines the limited feedback MIMO system model and the corresponding shape-gain product VQ structure. Section III derives the distortion measures and provides the optimal bit allocation solution; in general, proofs are deferred to appendices. This section also presents the linear precoding algorithm that incorporates the optimal bit allocation policy. Section IV presents results of numerical simulations illustrating the theory developed. The paper wraps up with some conclusions in Section V.

II. SYSTEM MODEL

We begin by developing the system model for linearly-precoded MU-MIMO system followed by the model for CSI feedback and the product shape-gain quantization structure.

A. MU MIMO System Model

Consider a single base station equipped with M transmit antennas communicating with K independent users. User k has N_k antennas and receives L_k data streams. All data streams are independent of each other. Let $L = \sum_k L_k$, $N = \sum_k N_k$. To ensure resolvability, we require $L \leq M$ and $L_k \leq N_k$.

Let $\mathbf{U} \in \mathcal{C}^{M \times L}$ denote the global precoder, the columns of which are unit-norm. Similarly, let $\mathbf{P} \in \mathcal{R}^{L \times L}$ denote the diagonal power matrix whose entries are the powers allocated to individual streams. Let, P_{\max} be the total available power; we require $\text{tr}[\mathbf{P}] \leq P_{\max}$. The data vector $\mathbf{x} = [x_1, \dots, x_L]^T = [\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_K^T]^T$, includes all L data streams to the K users. The $N_k \times M$ block fading channel, \mathbf{H}_k^H , between the BS and user k is assumed to be flat. The global channel matrix is \mathbf{H}^H , with $\mathbf{H} = [\mathbf{H}_1, \dots, \mathbf{H}_K]$. The elements of channel entries are assumed to be zero mean complex Gaussian random variables with unit variance. User k receives

$$\mathbf{y}_k^{DL} = \mathbf{H}_k^H \mathbf{U} \sqrt{\mathbf{P}} \mathbf{x} + \mathbf{n}_k, \quad (1)$$

where \mathbf{n}_k represents the zero mean additive white Gaussian noise at the receiver. User k , in order to estimate its own transmitted symbols from \mathbf{y}_k^{DL} , forms

$$\hat{\mathbf{x}}_k = \mathbf{\Lambda}_k \mathbf{V}_k^H \mathbf{y}_k^{DL}, \quad (2)$$

where $\mathbf{V}_k \mathbf{\Lambda}_k$ is the $N_k \times L_k$ decoder matrix for user k . The columns of $\mathbf{V}_k \in \mathcal{C}^{N_k \times L_k}$ are unit norm while $\mathbf{\Lambda}_k = \text{diag}(\lambda_{k1}, \lambda_{k2}, \dots, \lambda_{kL_k}) \in \mathcal{R}^{L_k \times L_k}$ contains the gain variables that normalize

the received data. Although the gain variables at the receiver side do not affect the signal-to-interference-plus-noise ratio (SINR), they play an important role in the error performance of transmissions that include amplitude modulation, e.g., quadrature amplitude modulated systems. Figure 1 illustrates the proposed downlink system.

Let $\mathbf{V}\mathbf{\Lambda}$ be the $N \times L$ block diagonal global decoder matrix, $\mathbf{V} = \text{diag}(\mathbf{V}_1, \dots, \mathbf{V}_K) \in \mathcal{C}^{N \times L}$ and $\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}_1, \dots, \mathbf{\Lambda}_K) \in \mathcal{R}^{L \times L}$. Overall,

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{\Lambda}\mathbf{V}^H\mathbf{H}^H\mathbf{U}\sqrt{\mathbf{P}}\mathbf{x} + \mathbf{V}^H\mathbf{n} \\ &= \mathbf{F}^H\mathbf{U}\sqrt{\mathbf{P}}\mathbf{x} + \mathbf{V}^H\mathbf{n},\end{aligned}\tag{3}$$

where, $\mathbf{n} = [\mathbf{n}_1^T, \mathbf{n}_2^T, \dots, \mathbf{n}_K^T]^T$. For the ease of representation, we define the $M \times L$ matrix $\mathbf{F} = \mathbf{H}\mathbf{V}\mathbf{\Lambda}$ with $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_L]$. The vectors $\mathbf{f}_1, \dots, \mathbf{f}_L$ are the effective $M \times 1$ vector downlink channels of the individual data streams.

The MSE of the i^{th} data stream of the k^{th} user is given by¹,

$$e_{k,i}^{DL} = E \left[(\hat{x}_{k,i} - x_{k,i}) (\hat{x}_{k,i} - x_{k,i})^H \right]. \tag{4}$$

The min-SMSE optimization problem is:

$$\min_{\mathbf{p}, \mathbf{U}, \mathbf{V}, \mathbf{\Lambda}} \sum_{k=1}^K \sum_{i=1}^{L_k} e_{k,i}^{DL}; \quad \text{subject to } \text{tr}[\mathbf{P}] \leq P_{\max}, ||\mathbf{u}_\ell|| = ||\mathbf{v}_\ell|| = 1, \tag{5}$$

To solve this problem, it is computationally efficient to use a virtual dual uplink [4]. In this uplink the transmit powers are $\mathbf{Q} = \text{diag}[q_1, \dots, q_L]^T$ for the L data streams, while the matrices \mathbf{U} and \mathbf{V} remain the same as before.

B. Feedback Model:

As mentioned earlier, we use an eigen-mode strategy [1], [2]. According to this strategy, the k^{th} user estimates its own channel \mathbf{H}_k and uses a set of dominant singular values and singular vectors of \mathbf{H}_k as $\mathbf{\Lambda}_k$ and \mathbf{V}_k respectively.

Since the user is aware of \mathbf{H}_k , \mathbf{V}_k and $\mathbf{\Lambda}_k$, it can form the product matrix $\mathbf{F}_k = \mathbf{H}_k\mathbf{V}_k\mathbf{\Lambda}_k$, whose columns act as the effective vector downlink channels for the data streams. Each user quantizes its effective vector downlink channel based on an Euclidean distance measure and feeds

¹Note that we, interchangeably, index streams as being the ℓ^{th} of L streams overall or the i^{th} stream of the k^{th} user. Any one-to-one mapping between the two notations is acceptable.

back the quantized channel to the BS. Details of the CSI quantization policy will be described in the next section. To model the effect of quantization, we consider the following relation between the original and the quantized variables,

$$\mathbf{f}_{k,i} = \hat{\mathbf{f}}_{k,i} + \tilde{\mathbf{f}}_{k,i} \text{ or } \mathbf{F} = \hat{\mathbf{F}} + \tilde{\mathbf{F}}. \quad (6)$$

Here, $\mathbf{f}_{k,i}$ denotes the effective vector downlink channel of the i^{th} stream of the k^{th} user. \mathbf{F} comprises L effective channel vectors with the original channel directions and channel gains. $\hat{\mathbf{F}}$ denotes the L quantized feedback vectors. The matrix $\tilde{\mathbf{F}}$ represents the quantization error.

The BS assumes that the quantization error matrix $\tilde{\mathbf{F}}$ has $M \times L$ independent identically Gaussian distributed (i.i.d.) elements with zero mean and a variance of σ_E^2/M , where σ_E^2 is the quantization error variance associated with each quantized vector $\hat{\mathbf{f}}_{k,i}$ and is defined as,

$$\sigma_E^2 = E \left[\|\mathbf{f}_{k,i} - \hat{\mathbf{f}}_{k,i}\|^2 \right]. \quad (7)$$

By using the optimal \mathbf{P} and \mathbf{U} , the minimum SMSE takes the following form [19]:

$$SMSE = L - M + \left(\sigma^2 + \frac{\sigma_E^2}{M} P_{\max} \right) \text{tr} [\mathbf{J}^{-1}], \quad (8)$$

where,

$$\mathbf{J} = \hat{\mathbf{F}} \mathbf{Q} \hat{\mathbf{F}}^H + \left(\sigma^2 + \frac{\sigma_E^2}{M} P_{\max} \right) \mathbf{I}_M. \quad (9)$$

where \mathbf{Q} is the virtual uplink power allocation matrix.

Equations in (8) and (9) show that the SMSE is directly related to the quantization error σ_E^2 . The limited feedback system design problem can therefore be formulated as minimization of the quantization error variance subject to a fixed feedback overhead.

C. Shape-Gain Product Quantization Model

We intend to find the optimal bit allocation for quantizing the effective vector downlink channel $\mathbf{f}_{k,i}$. From now on, we will use \mathbf{z} to represent the effective vector downlink channel to simplify the notation. According to the eigen-based receiver structure assumed in this work, \mathbf{z} represents the product of a singular value of the channel matrix and its corresponding singular vector.

Let $\hat{\mathbf{z}}$ be the quantized effective vector downlink channel and let $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{N_{tot}}\}$ denote the codebook of quantized channels. Here, $N_{tot} = 2^B$ are the total number of quantization levels using a total of B bits. This codebook is simplified to a product codebook. Fig. 2 illustrates the

product codebook operation based on independent quantization of gain and shape. Let, $\mathbf{z} = g\mathbf{s}$ where, $g = \|\mathbf{z}\|_2$ and the unit-norm $\mathbf{s} = \mathbf{z}/\|\mathbf{z}\|_2$ denote the gain and shape of the channel respectively. The BS is provided with the quantized information $\hat{\mathbf{z}} = \hat{g}\hat{\mathbf{s}}$, where \hat{g} and $\hat{\mathbf{s}}$ denote the quantized gain and shape respectively.

Let B_g and B_s denote the number of bits allocated to gain and shape quantization and define $N_g = 2^{B_g}$ and $N_s = 2^{B_s}$. Further, let \mathcal{C}_g and \mathcal{C}_s represent the gain and shape codebook respectively:

$$\mathcal{C}_g = [c_{g1}, c_{g2}, \dots, c_{gN_g}] \quad (10)$$

$$\mathcal{C}_s = [\mathbf{c}_{s1}, \mathbf{c}_{s2}, \dots, \mathbf{c}_{sN_s}]. \quad (11)$$

The product codebook can therefore be represented as,

$$\mathcal{C} = \mathcal{C}_g \times \mathcal{C}_s. \quad (12)$$

The quantized gain and shape variables are computed as:

$$\hat{g} = \arg \min_{c_g \in \mathcal{C}_g} (g - c_g)^2 \quad (13)$$

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{c}_s \in \mathcal{C}_s} \|\mathbf{s} - \mathbf{c}_s\|^2. \quad (14)$$

The Lloyd-Max algorithm is the optimal solution to find the codebook for the gain of the channel vector with the MSE objective [20]. We use the K -means approach, as described in [21], for numerical implementation of the Lloyd-Max algorithm. The optimal codebook of unit norm vectors with a Euclidean measure is not yet known. Therefore, we adopt random VQ to find the shape codebook. With this approach, the unit norm quantized shape vectors are randomly and independently distributed on the complex unit hyper-sphere in \mathbb{C}^M .

The remaining question is, given B , what is the optimal choice of B_s and B_g ?

III. DISTORTION ANALYSIS AND OPTIMAL BIT ALLOCATION SOLUTION

A. Design Objective

Our main problem is to optimize the shape-gain bit allocation as formulated below,

$$[B_s^*, B_g^*] = \arg \min_{B_s, B_g} E [\|\mathbf{z} - \hat{g}\hat{\mathbf{s}}\|^2] \quad (15)$$

$$\text{subject to : } B_s + B_g = B, B_s \geq 0, B_g \geq 0, \hat{g} \in \mathcal{C}_g, \hat{\mathbf{s}} \in \mathcal{C}_s.$$

Hamkins et al. [14] have shown that, for high resolution quantization (large B_s and B_g), the distortion measure takes the following form [14]:

$$E[||\mathbf{z} - \hat{g}\hat{\mathbf{s}}||^2] \approx E[(g - \hat{g})^2] + E[g^2] E[||\mathbf{s} - \hat{\mathbf{s}}||^2] \quad (16)$$

$$\approx D_g + E[g^2] D_s, \quad (17)$$

where, $E[g^2]$ denotes the variance of the gain and $D_g = E[(g - \hat{g})^2]$ is the gain quantization distortion. On the other hand, $D_s = E[||\mathbf{s} - \hat{\mathbf{s}}||^2]$ represents the distortion due to unit-norm shape quantization. Since D_g and D_s are independent of each other in (17), the optimal bit allocation problem can be solved using the following three steps:

- 1) Find D_g , gain distortion, for a given B_g .
- 2) Find D_s , shape distortion, for a given B_s .
- 3) Provide optimal bit allocation to minimize the overall distortion, i.e., $E[g^2] D_s + D_g$.

B. Distortion due to Gain Quantization

The distortion due to quantizing the gain is given by

$$D_g = E[(g - \hat{g})^2] = \int_0^\infty (r - \hat{g}(r))^2 f_g(r) dr. \quad (18)$$

Here, $\hat{g}(r)$ is the quantized value of r and $f_g(r)$ is the probability density function (pdf) of the gain. Using Bennett's integral ([17], page-186), the distortion in (18) takes the form,

$$D_g = \frac{1}{12N_g^2} ||f_g(r)||_{\frac{1}{3}}, \quad (19)$$

where, $N_g = 2^{B_g}$ and

$$||f_g(r)||_{\frac{1}{3}} = \left(\int_0^\infty |f_g(r)|^{\frac{1}{3}} dr \right)^3. \quad (20)$$

Lemma 1: For Rayleigh fading and based on the pdf of the dominant eigenvalues of Wishart matrix and Jacobian transformation in [22] and [23], we have,

$$||f_g(r)||_{\frac{1}{3}} = \frac{3 \times 3^{L(e)} \beta}{4(L(e) - 1)!} \Gamma^3 \left(\frac{L(e) + 1}{3} \right), \quad (21)$$

where, $L(e) = (M - e)(N_k - e)$, M represents the total number of transmit antennas at the BS, N_k denotes the number of receiver antennas of the k^{th} user. e denotes the index of the ordered eigenvalues where 0 represents the most dominant one, 1 denotes the 2nd most dominant one and so on. Finally, $\beta = \tilde{\lambda}_e / L(e)$ where $\tilde{\lambda}_e$ is the mean of the e^{th} eigenvalue.

Proof: See Appendix A. ■

Using (19) and (21), the gain distortion at high resolution can be expressed as,

$$D_g = \frac{1}{12N_g^2} \|f_g(r)\|_{\frac{1}{3}} \quad (22)$$

$$= \frac{1}{16N_g^2} \frac{3^{L(e)}\beta}{(L(e)-1)!} \Gamma^3 \left(\frac{L(e)+1}{3} \right) \quad (23)$$

$$= K_g 2^{-2B_g}, \quad (24)$$

where, $K_g = \frac{1}{16} \frac{3^{L(e)}\beta}{(L(e)-1)!} \Gamma^3 \left(\frac{L(e)+1}{3} \right)$ is a constant with respect to B_g . Equation (24) suggests that the gain distortion due to quantization is proportional to 2^{-2B_g} .

Figure 3 shows the distortion due to gain quantization of the dominant singular value of a 2×2 MIMO channel. As the figure verifies, the analytical expression converges to the simulation result as B_g increases.

C. Shape Quantization Distortion

This section focuses on the shape quantization distortion of a unit-norm vector in \mathbb{C}^M , in terms of the Euclidean distance. The Euclidean distance of two points in a \mathbb{C}^M plane has a one-to-one relation with the distance of two points in a \mathbb{R}^{2M} plane. Therefore, we can focus on quantization of unit-norm vectors in \mathbb{R}^{2M} instead of \mathbb{C}^M .

Figure 4 shows a two dimensional view of the problem where $OB = \mathbf{s}$, $OA = \hat{\mathbf{s}}$. Here, $\|\mathbf{s}\|_2 = \|\hat{\mathbf{s}}\|_2 = 1$. The Euclidean distance between \mathbf{s} and $\hat{\mathbf{s}}$ is defined by, $d = \|\mathbf{s} - \hat{\mathbf{s}}\|_2$. Define \mathcal{U}_{2M} as the unit hypersphere in \mathbb{R}^{2M} . The surface area of \mathcal{U}_{2M} is given by [15]

$$S(\mathcal{U}_{2M}) = 2MC_{2M}, \quad (25)$$

where,

$$C_{2M} = \frac{\pi^M}{\Gamma(M+1)}. \quad (26)$$

Define the spherical cap \mathcal{D} , i.e., the region ABC around \mathbf{s} in Fig. 4, as:

$$\mathcal{D} = (\hat{\mathbf{s}} \in \mathcal{U}_{2M} | \|\mathbf{s} - \hat{\mathbf{s}}\|_2 \leq d), \quad (27)$$

and let $\angle AOB = \theta$ be the angular distance between \mathbf{s} and $\hat{\mathbf{s}}$. Since $\|OA\|_2 = \|\hat{\mathbf{s}}\|_2 = 1$, we have $AD = \sin(\theta)$ and $OD = \cos(\theta)$. Also, since $\|OB\|_2 = \|\mathbf{s}\|_2 = 1$, we have $BD = 1 - \cos(\theta)$. Therefore,

$$AB^2 = AD^2 + BD^2 = \sin^2(\theta) + (1 - \cos(\theta))^2 = 2 - 2\cos(\theta). \quad (28)$$

Here, if we define $b = d^2$, we will have:

$$\theta = \cos^{-1}(1 - 0.5b). \quad (29)$$

The surface area of \mathcal{D} is given by [15],

$$S(\mathcal{D}) = (2M - 1)C_{2M-1} \int_0^\theta \sin^{2M-2} \phi d\phi. \quad (30)$$

Now, if we assume a small spherical cap of radius d centered on \mathbf{s} , the quantized vector can lie anywhere on this cap. Hence,

$$Pr[||\mathbf{s} - \hat{\mathbf{s}}||^2 \leq b] = \frac{S(\mathcal{D})}{S(\mathcal{U}_{2M})}. \quad (31)$$

Using (25), (26), (29) and (30) in (31) we get

$$Pr[||\mathbf{s} - \hat{\mathbf{s}}||^2 \leq b] = \frac{(2M - 1)C_{2M-1} \int_0^{\cos^{-1}(1-0.5b)} \sin^{2M-2} \phi d\phi}{2MC_{2M}}. \quad (32)$$

Since all the quantized vectors are randomly chosen, the probabilities that the square of the Euclidean distance between any vector in the codebook and the corresponding channel is higher than b , are independent of each other. Therefore,

$$Pr[\min_{i \in [1, N_s]} ||\mathbf{s} - \hat{\mathbf{s}}_i||^2 \geq b] = \left(1 - \frac{(2M - 1)C_{2M-1} \int_0^{\cos^{-1}(1-0.5b)} \sin^{2M-2} \phi d\phi}{2MC_{2M}}\right)^N. \quad (33)$$

Hence, expected value of the distortion error due to shape quantization can be calculated as follows:

$$E(b) = \int_0^4 Pr[\min_{i \in N} ||\mathbf{s} - \hat{\mathbf{s}}_i||^2 \geq b] db. \quad (34)$$

The limits of integration in (34) follow from the fact that the square of the Euclidean distance between two points on a unit radius sphere has a range of 0 to 4.

Lemma 2:

$$E(b) < K_s 2^{\frac{-2Bs}{2M-1}}, \quad (35)$$

where,

$$K_s = \left(\frac{\pi^{\frac{2M-1}{2}} \Gamma(M)}{2\pi^M \Gamma(\frac{2M-1}{2} + 1)} \right)^{\frac{-2}{2M-1}}. \quad (36)$$

is a constant.

Proof: See Appendix B. ■

Figure 5 shows that the upper bound of the shape distortion in (35) has a fixed gap with respect to the simulation result. Therefore, we can safely approximate the shape distortion with the analytical expression in (37). Thus,

$$D_s = E (||\mathbf{s} - \hat{\mathbf{s}}||^2) \approx \left(\frac{\pi^{\frac{2M-1}{2}} \Gamma(M)}{2\pi^M \Gamma(\frac{2M-1}{2} + 1)} \right)^{\frac{-2}{2M-1}} 2^{\frac{-2B_s}{2M-1}} = K_s 2^{\frac{-2B_s}{2M-1}}. \quad (37)$$

D. Optimal Bit Allocation

Having analyzed the individual terms in (17), we are now able to answer the core question of this paper: the allocation of bits across gain and shape. In (17), the overall distortion measure was shown to take the following form,

$$D = E [||\mathbf{z} - \hat{g}\hat{\mathbf{s}}||^2] \approx D_g + E [g^2] D_s. \quad (38)$$

Using the gain and shape distortion measures of (24) and (37), D can be approximated as,

$$D \approx E [g^2] K_s 2^{-\frac{2B_s}{2M-1}} + K_g 2^{-2B_g} \quad (39)$$

$$\approx \bar{K}_s 2^{-\frac{2B_s}{2M-1}} + K_g 2^{-2(B-B_s)}, \quad (40)$$

where $\bar{K}_s = K_s E [g^2]$. With these relations in hand, the optimal shape-gain bit allocation can be formulated as follows,

$$B_s^* = \arg \min_{B_s} \bar{K}_s 2^{-\frac{2B_s}{2M-1}} + K_g 2^{-2(B-B_s)} \quad (41)$$

Theorem 1:

The optimal bit allocation problem has the following solution:

$$B_s = \frac{2M-1}{2M} B + \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right) \quad (42)$$

$$B_g = B - B_s = \frac{1}{2M} B - \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right). \quad (43)$$

Here, \bar{K}_s and K_g are the terms defined in the previous subsections.

Proof: See Appendix C. ■

Note that, \bar{K}_s and K_g in (42) and (43) depend on M but not B . Therefore, as B goes to infinity,

$$B_s \approx \frac{2M-1}{2M} B \quad (44)$$

$$B_g \approx \frac{1}{2M} B. \quad (45)$$

The analytical expressions of (44) and (45) can be intuitively explained as follows: The norm of a \mathbb{C}^M vector varies across a one dimensional line. However, the shape of a \mathbb{C}^M vector is uniformly distributed in the surface of a $(2M - 1)$ dimensional hypersphere. Therefore, given $2M$ number of bits to quantize a \mathbb{C}^M vector, one should expend approximately 1 and $(2M - 1)$ bits to quantize the gain and shape of the vector respectively. It is worth noting that, from a different point of view and using a very different analysis, the work in [15], [16] leads to a similar expression and explanation. However, this similarity is only for a high available feedback rate.

Finally, the overall quantization error for a fixed feedback overhead takes the following form:

$$D = \bar{K}_s 2^{-\frac{2B_s}{2M-1}} + K_g 2^{-2B_g} \quad (46)$$

$$\begin{aligned} &= \bar{K}_s 2^{-\frac{2}{2M-1} \left(\frac{2M-1}{2M} B + \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g (2M-1)} \right) \right)} + K_g 2^{-2 \left(\frac{1}{2M} B - \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g (2M-1)} \right) \right)} \\ &= 2^{-\frac{B}{M} \log_2 \left(\frac{\bar{K}_s}{K_g (2M-1)} \right)} \left(\bar{K}_s 2^{-\frac{1}{2M}} - K_g 2^{-\frac{2M-1}{2M}} \right) \end{aligned} \quad (47)$$

$$= D_c 2^{-\frac{B}{M}}, \quad (48)$$

where, $D_c = \log_2 \left(\frac{\bar{K}_s}{K_g (2M-1)} \right) \left(\bar{K}_s 2^{-\frac{1}{2M}} - K_g 2^{-\frac{2M-1}{2M}} \right)$ is a constant.

E. Overall Linear Precoding Algorithm

In the previous section we derived the optimal allocation of available bits across gain and shape. Here we use this information to develop the overall linear precoding algorithm. The material exploits previous work in [19], [24]. The algorithm steps are:

- 1) A gain codebook of B_g bits is generated based on the dominant singular values of a random Gaussian matrix and using the K -means algorithm [21]. A shape codebook of 2^{B_s} random unit-norm vectors, uniformly and independently distributed in \mathbb{C}^M , is also generated. Both these codebooks are generated off-line and the codebooks are shared between the BS and the users.
- 2) The BS sends common pilot symbols so that the receivers can estimate \mathbf{H}_k .
- 3) The receivers calculate the dominant singular values Λ_k and the corresponding singular vectors \mathbf{V}_k and form $\mathbf{F}_k = \mathbf{H}_k \mathbf{V}_k \Lambda_k$.
- 4) The receivers use the codebooks to quantize the gain and shape of the column vectors in \mathbf{F}_k and feedback the quantization indices to the BS.

- 5) The BS calculates the optimum virtual uplink power allocation matrix as,

$$\mathbf{Q}^{opt} = \min_{\mathbf{Q}} \left(\sigma^2 + \frac{\sigma_E^2 P_{\max}}{M} \right) \text{tr}(\mathbf{J}^{-1}), \text{ such that } \text{tr}(\mathbf{Q}) \leq P_{\max}.$$

Here, $\sigma_E^2 = D$ as in (48) and \mathbf{J} is calculated according to (9).

- 6) The precoding matrix of the k^{th} user is calculated as, $\mathbf{U}_k^{mmse} = \mathbf{J}^{-1} \hat{\mathbf{F}}_k \sqrt{\mathbf{q}_k}$. Here, $\mathbf{q}_k = [q_{k1}, \dots, q_{kL_k}]$ contains the virtual uplink power variables of the L_k streams of the k^{th} user.

- 7) Using a recent result [25], the downlink transmit power variables are determined as, $\mathbf{p} = \mathbf{q}$.

Here, $\mathbf{q} = [q_1, \dots, q_L]$ is the virtual uplink transmit powers of the L streams.

IV. NUMERICAL SIMULATIONS

This section provides the results of simulations to study the effect of shape-gain quantization on the performance of the MU MIMO linear precoding scheme described. We assume the following scenario in the simulation setup: the base station has two transmit antennas and serves two receivers in the downlink. Each receiver has 2 receive antennas and receives 1 data stream. The feedback overhead per user, B , is 16 bits. We show performance curves for different shape quantization bits, B_s , in Figs. 6, 7 and 8. The corresponding number of gain quantization bits is given by, $B_g = B - B_s$.

Figure 6 illustrates the effect of bit allocation on the quantization error and suggests that $B_s = 13$ and $B_g = 3$ are the optimal bit allocations for this scenario. The analytical results in (42) and (43) lead to, $B_g = 2.6$, $B_s = 13.4$ which matches the numerical result.

Fig. 7 plots the SMSE of the same system with the transmitter using 16-QAM. The figure shows that $B_s = 12$ leads to the minimum SMSE. The SMSE performance of $B_s = 13$, i.e., the optimal solution obtained from analytical results, is very close to that of $B_s = 12$ bits. The minor difference between the simulation and analytical result stems from the fact that the derived gain distortion holds only for large number of gain quantization bits. Note that, $B_s = 16$, i.e., quantizing the shape exclusively, leads to much higher SMSE. Therefore, optimal bit allocation across gain and shape feedback provides better performance in terms of SMSE.

Fig. 8 shows that the bit allocation $B_s = 12$ or $B_s = 13$ also lead to better performance in terms of BER. If one uses all the bits for direction quantization, i.e., $B_s = 16$, the effect of multiuser interference on the norm of the received signal cannot be rectified. This leads to the inferior performance of $B_s = 16$.

We did not plot the performance of all possible bit allocations, e.g., $B_s = 0$ to $B_s = 8$, so that the figures look clearer. However, the performance trend of $B_s = 11$ to $B_s = 9$ bit clearly suggests that full gain quantization ($B_s = 0$, $B_g = 16$) will also perform much inferior to the optimal bit allocation in shape-gain quantization. Thus, optimal shape-gain quantization can improve over full gain or full shape quantization and lead to a lower BER in multiuser MIMO systems. In wireless ethernet [26] systems, where a small number of bit errors may lead to the whole packet drop [27], optimal bit allocation in shape-gain quantization can significantly reduce the packet loss rate and save packet re-transmission time.

V. CONCLUSION

This paper studies the optimal bit allocation across gain and shape quantization in a MU MIMO downlink system by minimizing the SMSE of the system for a fixed feedback overhead per user. We show that the distortion due to gain and shape quantization are proportional to 2^{-2B_g} and $2^{-\frac{2B_s}{2M-1}}$ respectively, suggesting that, in the asymptotic region of high feedback overhead, the number of shape quantization bits should be approximately $(2M - 1)$ times than the number of gain quantization bits. The analysis and importance of bit allocation is borne out by the simulation results that show significant worse performance for the usual approach (in MU MIMO downlink systems) of only quantizing the gain or shape but not both.

Our work with respect to the gain distortion calculation is quite general, since the gain quantization distortion of other distributions like Rician, Nakagami and Weibull fading can also be calculated using Bennett's integral. However, the optimal bit allocation results might be different from the Rayleigh fading case considered in this paper.

APPENDIX A

PROOF OF LEMMA 1

The authors of [22] have provided the following pdf of the eigenvalues of a MIMO channel,

$$f(\lambda_e) = \frac{1}{(L(e) - 1)!} \frac{\lambda_e^{L(e)-1}}{\beta^{L(e)}} \exp\left(-\frac{\lambda_e}{\beta}\right). \quad (49)$$

Here, λ_e denotes the e^{th} eigenvalue of the Wishart matrix (i.e., $\mathbf{H}^H \mathbf{H}$ or $\mathbf{H} \mathbf{H}^H$). e denotes the index of the ordered eigenvalues. $L(e) = (M - e)(N_k - e)$. β is a constant whose value is given

through the following equation,

$$\beta = \frac{\tilde{\lambda}_e}{L(e)}. \quad (50)$$

Here $\tilde{\lambda}_e$ is the mean of the eigenvalue. (49) provides the probability distribution function of the eigenvalue of the Wishart matrix, λ_e . In our proposed algorithm, we are trying to quantize g , the singular values of the Gaussian matrix \mathbf{H} . Now, $\lambda_e = g^2$.

Using Jacobian transformation [23], the probability distribution of the singular values of the Gaussian matrix can be found as follows,

$$f_g(r) = \frac{1}{(L(e) - 1)!} \frac{(r^2)^{L(e)-1}}{\beta^{L(e)}} \exp\left(-\frac{r^2}{\beta}\right) 2r. \quad (51)$$

Therefore,

$$\|f_g(r)\|_{\frac{1}{3}} = \frac{2}{(L(e) - 1)!} \frac{1}{\beta^{L(e)}} \left(\int_0^\infty r^{\frac{2L(e)-1}{3}} \exp\left(-\frac{r^2}{3\beta}\right) dr \right)^3. \quad (52)$$

Using standard mathematical tables of [28] (P - 380, eqn - 662), we find

$$\int_0^\infty x^n \exp(-ax^p) dx = \frac{\Gamma\left(\frac{n+1}{p}\right)}{pa^{\left(\frac{n+1}{p}\right)}}. \quad (53)$$

Comparing (53) with (52), we find, $n = \frac{2L(e)-1}{3}$, $a = \frac{1}{3\beta}$, $p = 2$. Therefore,

$$\left(\int_0^\infty r^{\frac{2L(e)-1}{3}} \exp\left(-\frac{r^2}{3\beta}\right) dr \right) = \frac{\Gamma\left(\frac{\frac{2L(e)-1}{3}+1}{2}\right)}{2\left(\frac{1}{3\beta}\right)^{\frac{\frac{2L(e)-1}{3}+1}{2}}} \quad (54)$$

$$\left(\int_0^\infty r^{\frac{2L(e)-1}{3}} \exp\left(-\frac{r^2}{3\beta}\right) dr \right) = \frac{1}{2} (3\beta)^{\frac{L(e)+1}{3}} \Gamma\left(\frac{L(e)+1}{3}\right) \quad (55)$$

$$\left(\int_0^\infty r^{\frac{2L(e)-1}{3}} \exp\left(-\frac{r^2}{3\beta}\right) dr \right)^3 = \frac{1}{8} (3\beta)^{L(e)+1} \Gamma^3\left(\frac{L(e)+1}{3}\right). \quad (56)$$

Using (56) in (52), we get,

$$\|f_g(r)\|_{\frac{1}{3}} = \frac{2}{(L(e) - 1)!} \frac{1}{\beta^{L(e)}} \frac{1}{8} 3^{L(e)+1} \beta^{L(e)+1} \Gamma^3\left(\frac{L(e)+1}{3}\right) \quad (57)$$

$$= \frac{3 \times 3^{L(e)} \beta}{4(L(e) - 1)!} \Gamma^3\left(\frac{L(e)+1}{3}\right). \quad (58)$$

APPENDIX B
PROOF OF LEMMA 2

Using (33),

$$Pr[\min_{i \in N} \|\mathbf{s} - \hat{\mathbf{s}}_i\|^2 \geq b] = \left(1 - \frac{(2M-1)C_{2M-1} \int_0^{\cos^{-1}(1-0.5b)} \sin^{2M-2} \phi d\phi}{2MC_{2M}} \right)^N \quad (59)$$

$$= \left(1 - K_1 \int_0^{\cos^{-1}(1-0.5b)} \sin^{2M-2} \phi d\phi \right)^N \quad (60)$$

$$\approx \left(1 - K_1 \int_0^{\cos^{-1}(1-0.5b)} \phi^{2M-2} d\phi \right)^N \quad (61)$$

$$= \left(1 - K_2 (\cos^{-1}(1-0.5b))^{2M-1} \right)^N. \quad (62)$$

In (60), we assumed $K_1 = \frac{(2M-1)C_{2M-1}}{2MC_{2M}}$. (61) follows from the fact that, given a large number of quantization vectors, i.e., at high bit rate, the complementary cumulative distribution function (CCDF) is significant only for smaller values of ϕ . For these smaller angles, we can assume $\sin \phi \approx \phi$. Equation (62) follows from assuming $K_2 = \frac{K_1}{2M-1}$.

Figure 9 compares the simulated shape distortion with the original and approximate analytical shape distortion of a $2 \times 1 \mathbb{C}^M$ vector. Here, the CCDF of the original and approximate analytical expressions are superimposed with the simulated CCDF. Hence, (60) and (61) accurately model the actual distortion. This justifies the transition from (60) to (61).

Now, using (34), we find,

$$E(b) = \int_0^4 Pr[\min_{i \in N} \|\mathbf{s} - \hat{\mathbf{s}}_i\|^2 \geq b] db \quad (63)$$

$$= \int_0^a \left(1 - K_2 (\cos^{-1}(1-0.5b))^{2M-1} \right)^N db \quad (64)$$

$$= 2 \int_0^\psi (1 - K_2 \theta^{2M-1})^N \sin(\theta) d\theta \quad (65)$$

$$\approx 2 \int_0^\psi (1 - K_2 \theta^{2M-1})^N \theta d\theta \quad (66)$$

$$\approx 2 \int_0^1 (1 - K_2 \theta^{2M-1})^N \theta d\theta \quad (67)$$

$$= 2 \int_0^1 \left(\sum_{i=0}^N \binom{N}{i} (-1)^i K_2^i \theta^{i(2M-1)+1} \right) d\theta \quad (68)$$

$$= 2 \sum_{i=0}^N \frac{\binom{N}{i} (-1)^i K_2^i}{i(2M-1) + 2}. \quad (69)$$

The transition from (63) to (64) can be explained as follows: the similarity between (60) and (61) holds only for smaller values of b since $\sin \phi \neq \phi$ for larger ϕ . Therefore, although the square of the Euclidean distance between two random unit norm vectors can vary from 0 to 4, (62) holds only for a smaller range of b . At the presence of a large number of codewords, the squared distance between the original and the quantized channel takes large values with a negligibly small probability. Therefore, we can truncate the range of b as long as the CCDF of the original function is negligible outside the range, i.e., the limited range of b does not have any significant affect on the calculation of the expected value of the distortion. Using this analysis, in (64), we use a as the truncated range, i.e., we assume that b can vary from 0 to a .

In (65) we assumed, $\theta = (\cos^{-1}(1 - 0.5b))$. Therefore, $\psi = (\cos^{-1}(1 - 0.5a))$. Since only smaller angles of θ contribute to $E(b)$, we assumed $\sin \theta \approx \theta$ in (66). In (68), we assumed $\psi = 1$ to simplify the other calculations.

Fig. 10 justifies the approximations that we used in the derivations of shape distortion calculation. Here, approx1 and approx2 denote $\sin(\theta) \approx \theta$ (ref: eq. 66) and $\psi \approx 1$ (ref: eq. 67) respectively. As Fig. 10 shows, the three curves are superimposed with each other. Therefore, our justifications are valid for high bit rate quantization.

Applying $\binom{N}{i} = \frac{(-1)^i (-N)_i}{i!}$, where $(-N)_i = \frac{\Gamma(-N+i)}{\Gamma(-N)}$ [29], (69) takes the following form,

$$\sum_{i=0}^N \frac{(-1)^i (-N)_i (-1)^i K_2^i}{i!(i(2M-1) + 2)} = \frac{2}{2M-1} \sum_{i=0}^N \frac{(-N)_i K_2^i}{i!(i + \frac{2}{2M-1})} \quad (70)$$

$$= \frac{2}{2M-1} \frac{N!}{\frac{2}{2M-1} \left(1 + \frac{2}{2M-1}\right)_N} K_2^{\frac{-2}{2M-1}} \quad (71)$$

$$= \frac{N! \Gamma\left(1 + \frac{2}{2M-1}\right)}{\Gamma\left(N + 1 + \frac{2}{2M-1}\right)} K_3 \quad (72)$$

$$= \frac{N \Gamma(N) \Gamma\left(\frac{2M+1}{2M-1}\right)}{\Gamma\left(N + \frac{2M+1}{2M-1}\right)} K_3 \quad (73)$$

$$= N \beta\left(N, \frac{2M+1}{2M-1}\right) K_3. \quad (74)$$

(71) was found using ([30], 6.6.8). In (72), we assumed $K_3 = K_2^{\frac{-2}{2M-1}}$. (74) was obtained using the relation between the gamma and beta function, $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ [31]. Following a similar

work in [3], we find,

$$N\beta \left(N, \frac{2M+1}{2M-1} \right) = 2^B \frac{\Gamma(2^B) \Gamma(1 + \frac{2}{2M-1})}{\Gamma(2^B + 1 + \frac{2}{2M-1})} \quad (75)$$

$$\leq 2^B \frac{\Gamma(2^B)}{\Gamma(2^B + 1 + \frac{2}{2M-1})} \quad (76)$$

$$= \frac{\Gamma(2^B + 1)}{\Gamma(2^B + 1 + \frac{2}{2M-1})}. \quad (77)$$

The preceding inequality in (76) is justified by the following reasoning: due to the convexity of the gamma function [3] and the fact that $\Gamma(1) = \Gamma(2) = 1$, $\Gamma(x) \leq 1$ for $1 \leq x \leq 2$. Let, $y = 2^B + \frac{2}{2M-1}$, $t = 1 - \frac{2}{2M-1}$, so that, $y + t = 2^B + 1$, $y + 1 = 2^B + 1 + \frac{2}{2M-1}$. By applying Kershaw's inequality for the gamma function [32],

$$\frac{\Gamma(y+t)}{\Gamma(y+1)} < \left(y + \frac{t}{2} \right)^{t-1} \quad \forall y > 0, 0 < t < 1. \quad (78)$$

Using (78),

$$\frac{\Gamma(2^B + 1)}{\Gamma(2^B + 1 + \frac{2}{2M-1})} < \left(2^B + \frac{2}{2M-1} + 0.5 - \frac{1}{2M-1} \right)^{\frac{-2}{2M-1}} \quad (79)$$

$$= \left(2^B + \frac{1}{2M-1} + 0.5 \right)^{\frac{-2}{2M-1}} \quad (80)$$

$$< 2^{\frac{-2B}{2M-1}}. \quad (81)$$

Using (81) and the value of K_3 we find,

$$2 \sum_{i=0}^N \frac{(-1)^i (-N)_i (-1)^i K_2^i}{i! (i(2M-1) + 2)} < \left(\frac{C_{2M-1}}{2MC_{2M}} \right)^{-\frac{2}{2M-1}} 2^{\frac{-2B}{2M-1}}. \quad (82)$$

Using the values of C_{2M-1} and C_{2M} one can obtain,

$$E(b) < K_s 2^{\frac{-2B_s}{2M-1}}, \quad (83)$$

where, $K_s = \left(\frac{\pi^{\frac{2M-1}{2}} \Gamma(M)}{2\pi^M \Gamma(\frac{2M-1}{2} + 1)} \right)^{\frac{-2}{2M-1}}$ is a constant with respect to B_s .

APPENDIX C

PROOF OF THEOREM 1

Taking the 1st and 2nd order derivatives of (40), we find,

$$\frac{dD}{dB_s} = \bar{K}_s (\ln 2) 2^{-\frac{2B_s}{2M-1}} \left(-\frac{2}{2M-1} \right) + K_g (\ln 2) (2^{-2(B-B_s)}) 2 \quad (84)$$

$$\frac{d^2 D}{d^2 B_s} = \bar{K}_s (\ln 2)^2 2^{-\frac{2B_s}{2M-1}} \left(-\frac{2}{2M-1} \right)^2 + K_g (2 \ln 2)^2 (2^{-2(B-B_s)}). \quad (85)$$

From (85), $\frac{d^2 D}{d^2 B_s} \geq 0$. Therefore, the optimal bit allocation problem is convex [33]. Now, equating the 1st derivative to be zero,

$$\frac{\bar{K}_s}{2M-1} 2^{\frac{-2B_s}{2M-1}} = K_g 2^{-2(B-B_s)} \quad (86)$$

$$2^{-2B+2B_s+\frac{2B_s}{2M-1}} = \frac{\bar{K}_s}{K_g(2M-1)} \quad (87)$$

$$\frac{2MB_s}{2M-1} = B + \frac{1}{2} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right) \quad (88)$$

$$B_s = \frac{2M-1}{2M} B + \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right). \quad (89)$$

Therefore, at the optimal point,

$$B_s = \frac{2M-1}{2M} B + \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right) \quad (90)$$

$$B_g = \frac{1}{2M} B - \frac{2M-1}{4M} \log_2 \left(\frac{\bar{K}_s}{K_g(2M-1)} \right). \quad (91)$$

REFERENCES

- [1] F. Boccardi, H. Huang, and M. Trivellato, "Multiuser eigenmode transmission for MIMO broadcast channels with limited feedback," in *Proc. IEEE SPAWC 2007*, June 2007.
- [2] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Transactions on Signal Processing*, vol. 55, pp. 3417–3428, 2007.
- [3] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Communications*, vol. 52, pp. 5045–5060, Nov. 2006.
- [4] A. M. Khachan, A. J. Tenenbaum, and R. S. Adve, "Linear processing for the downlink in multiuser MIMO systems with multiple data streams," in *Proc. IEEE ICC'2006*, June 2006, vol. 9, pp. 4113–4118.
- [5] A. D. Dabbagh and D. J. Love, "Multiple antenna MMSE based downlink precoding with quantized feedback or channel mismatch," *IEEE Transactions on Communications*, vol. 7, pp. 1859–1868, Nov. 2008.
- [6] M. N. Islam and R. S. Adve, "SMSE precoder design in a multiuser miso system with limited feedback," in *Proc. Queen's Biennial Symposium on Communications 2010*, May 2010, pp. 352–356.
- [7] T. Kim and M. Skoglund, "Diversity-multiplexing tradeoff in mimo channels with partial csit," *IEEE Trans. Inf. Theory*, vol. 53, pp. 2743–2759, Aug. 2007.
- [8] S. Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. Communications*, vol. 50, pp. 785–798, May 2002.
- [9] A. Khoshnevis and A. Sabharwal, "On the asymptotic performance of multiple antenna channels with quantized feedback," *IEEE Trans. Wireless Comm.*, vol. 7, pp. 3869–3877, Oct. 2008.
- [10] A. Narula, M. Lopez, M. Trott, and G. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 1423–1436, Oct. 1998.
- [11] K. Mulkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2562–2579, Oct. 2003.

- [12] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Select. Areas Commun.*, vol. 25, pp. 1478–1491, Sept. 2007.
- [13] V. Lau, Y. Lau, and T. Chen, "On the design of MIMO block-fading channels with feedback-link capacity constraint," *IEEE Transactions on Communications*, vol. 52, no. 1, pp. 62–70, 2004.
- [14] J. Hamkins and K. Zeger, "Gaussian source coding with spherical codes," *IEEE Transactions on Information Theory*, vol. 48, pp. 2980–2989, Nov. 2002.
- [15] B. Khoshnevis and W. Yu, "Bit allocation law for multiantenna channel feedback quantization: Single-user case," *IEEE Transactions on Signal Processing*, vol. 59, pp. 2270–2283, May 2011.
- [16] B. Khoshnevis and W. Yu, "Bit allocation laws for multi-antenna channel quantization: Multi-user case," *IEEE Transactions on Signal Processing*, accepted, arXiv:1003.2259v2 [cs.IT].
- [17] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, MA, 1991.
- [18] S. Shi and M. Schubert, "MMSE transmit optimization for multi-user multi-antenna systems," in *Proc. IEEE ICASSP'2005*, Mar. 2008, vol. 3, pp. 409–412.
- [19] M. N. Islam and R. S. Adve, "Transceiver design using linear precoding in a multiuser system with limited feedback," *IET Journals on Communications*, vol. 5, pp. 27–38, Jan. 2011.
- [20] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, Mar. 1982.
- [21] Mathworks, "The mathworks website," <http://www.mathworks.com/access/helpdesk/help/toolbox/stats/kmeans.html>, 2010.
- [22] T. Taniguchi, S. Sha, and Y. Karasawa, "Analysis and approximation of statistical distribution of eigenvalues in i.i.d. MIMO channels under Rayleigh fading," *IEICE Trans. on Info. Th. and Its Appl.*, vol. E91-A, pp. 2808–2817, Oct. 2008.
- [23] G. Strang, *Linear Algebra & Its Applications*, Harcourt Brace Jovanovich Publishers, CA, 1988.
- [24] M. N. Islam and R. S. Adve, "Linear transceiver design in a multiuser MIMO system with quantized channel state information," in *Proc. IEEE ICASSP 2010*, Mar. 2010, pp. 3410–3413.
- [25] A. J. Tenenbaum and R. S. Adve, "Minimizing sum-MSE implies identical downlink and dual uplink power allocations," *IEEE Transactions on Communications*, vol. 59, pp. 686–688, Mar. 2011.
- [26] "IEEE standard 802.11-2008. part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," 2008.
- [27] D. G. Yoon, S. Y. Shin, W. H. Kwon, and H. S. Park, "Packet error rate analysis of ieee 802.11b under ieee 802.15.4 interference," in *Proc. IEEE VTC 2006-Spring*, May 2006, pp. 1186–1190.
- [28] W. H. Beyer, *CRC Standard Mathematical Tables, 26th Ed.*, CRC, FL, 1981.
- [29] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a miso system," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458–462, 2007.
- [30] E. R. Hansen, *A Table of Series & Products*, Prentice Hall, Inc., NJ, 1975.
- [31] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Process*, McGraw-Hill Companies, Inc., NY, 2002.
- [32] D. Kershaw, "Some extensions of the W. Gautschi's inequalities for the gamma function," *Mathematics of Computation*, vol. 41, no. 164, pp. 607–611, Oct. 1983.
- [33] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University, Cambridge, UK, 2004.

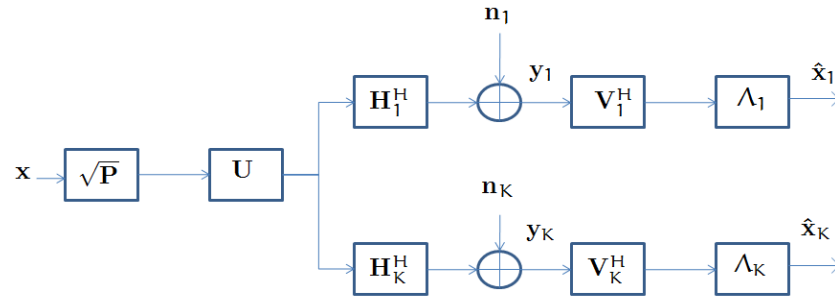


Fig. 1. MU MIMO system model in the downlink

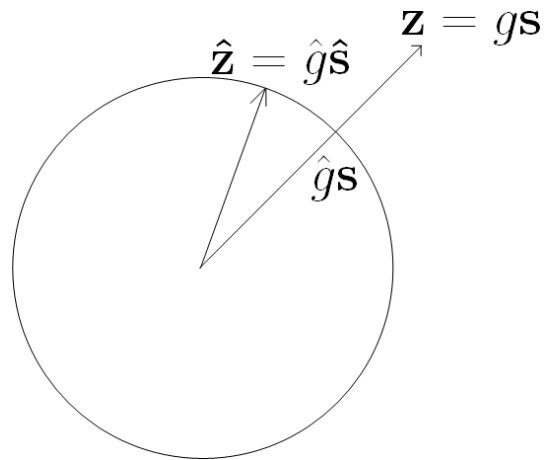


Fig. 2. Gain-shape product quantization

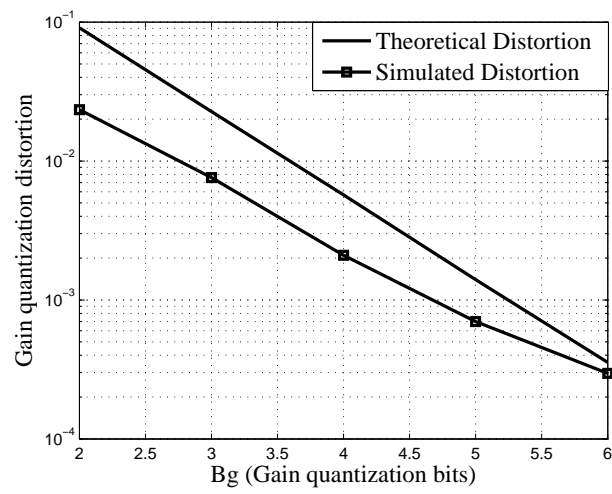


Fig. 3. Quantization distortion of the dominant singular value of 2x2 MIMO channel

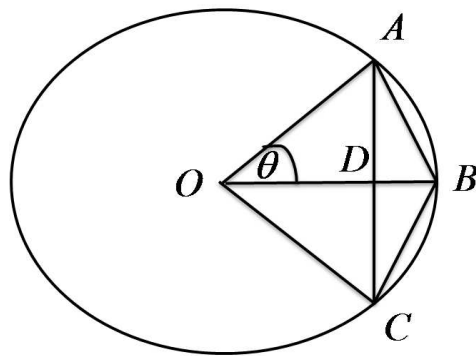


Fig. 4. Shape quantization block diagram

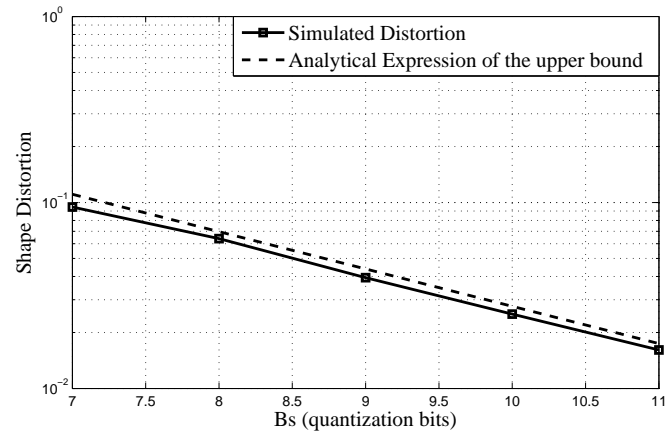


Fig. 5. Comparison of the simulated distortion with the theoretical upper bound (2x1 complex vector)

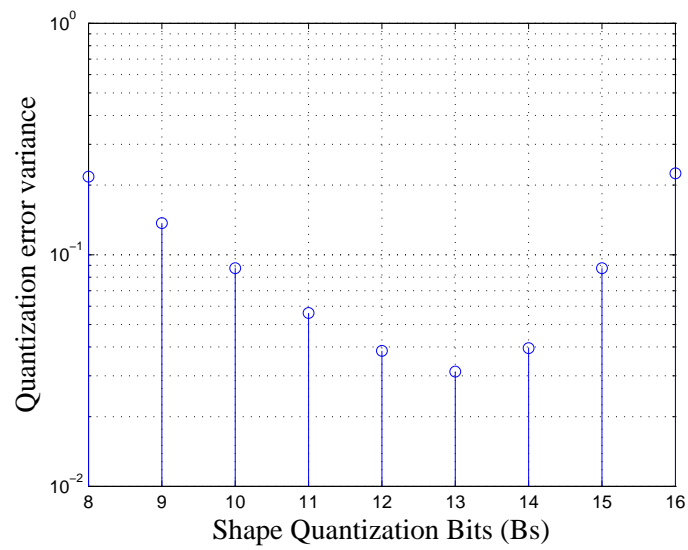


Fig. 6. Effect of bit allocation in the quantization of the product of dominant eigenvalue & the corresponding eigenvector of a 2 x 2 MIMO channel

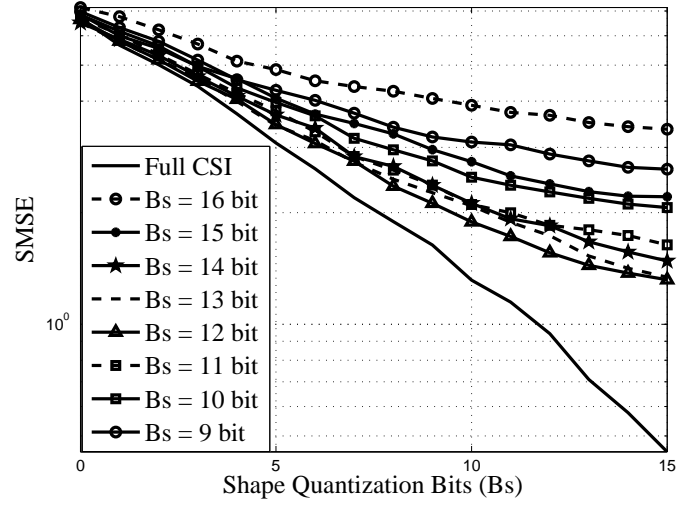


Fig. 7. Effect of bit allocation in the SMSE of 16-QAM system, $M = 2$, $N = [2 \ 2]$, $L = [1 \ 1]$, $B = 16$

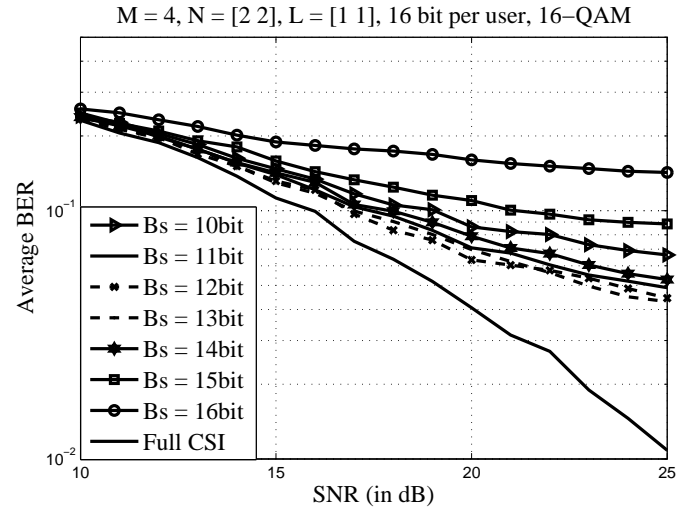


Fig. 8. Effect of bit allocation in the BER of 16-QAM systems, $M = 2$, $N = [2 \ 2]$, $L = [1 \ 1]$, $B = 16$

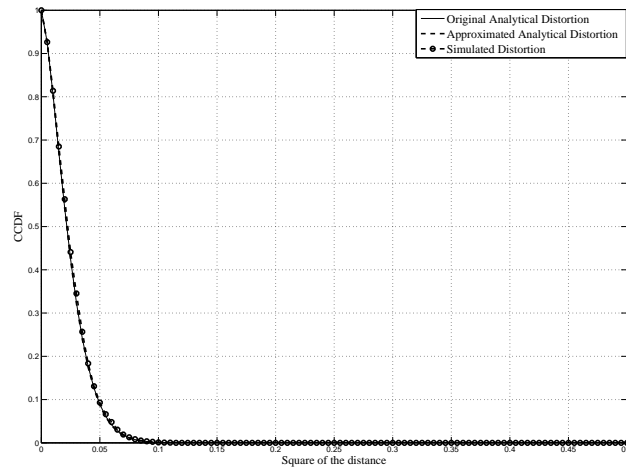


Fig. 9. Comparison of the original and approximated complementary cumulative distribution function of the shape distortion of a 2x1 vector (10 bit quantization)

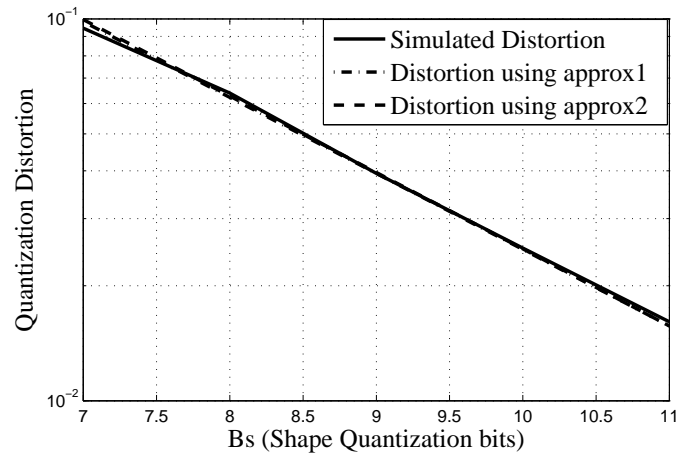


Fig. 10. Justification of the approximations used in Shape distortion calculation